Spring 2017

Quantifying the Effect of The Shift in Major League Baseball

Christopher John Hawke Jr.
Bard College, ch7749@bard.edu

Quantifying The Effect of The Shift in Major League Baseball

A Senior Project submitted to
The Division of Science, Mathematics, and Computing
of
Bard College

by
Jack Hawke

Annandale-on-Hudson, New York
May, 2017

Abstract

Baseball is a very strategic and abstract game, but the baseball world is strangely obsessed with statistics. Modern mainstream statisticians often study offensive data, such as batting average or on-base percentage, in order to evaluate player performance. However, this project observes the game from the opposite perspective: the defensive side of the game. In hopes of analyzing the game from a more concrete perspective, countless mathemeticians - most famously, Bill James - have developed numerous statistical models based on real life data of Major League Baseball (MLB) players. Large numbers of metrics go into these models, but what this project attempts to quantify specifically how positioning of defensive players in baseball ultimately affects the success of those players. This consults the use of multivariate probability distributions and models that allow us to study how good players really are based on data.

Contents

Dedication

I dedicate this project to my Mom and Dad, without whom I would not have the knowledge or resources that ultimately made this project possible.

Acknowledgments

A big thank you to Dr. Stefan Mendez-Diez, who played a tremendous role, both as my adviser in steering this project in the right direction, and also as an educator with respect to probability as a field of study, without which this project would not have been possible.

Another thank you to Dr. Shane Jensen, Dr. Kenneth Shirley, and Dr. Abraham Wyner, who's outstanding article published in *The Annals of Applied Statistics* serves as the basis for this project.

Thanks also to the faculty of the Department of Mathematics of Bard College, and to all my math teachers throughout my life.

Thank you so much.

1
Introduction

After seeing the movie *Moneyball*, I was amazed by the fact that mathematics can directly impact and improve decisions made both on and off the baseball field. Going into this project, I knew I wanted to somehow utilize sabermetrics to study and present an idea, or series of them, but did not know much of the specific methods and actual math that went into these models that Bill James helped to create.

I knew that I could not simply provide proofs and examples of the math that people have already been doing for decades. Clearly I would need to utilize models that have already been built, but I knew I wanted to craft my own question. Upon graduation, I will have lettered four years as a member and three years as a captain of the Bard baseball team. I have been playing, watching, teaching, and studying the game for many years, and although I consider myself an expert on baseball strategy, I certainly am not an expert on sabermetrics or the mathematical models that can and have represent it.

We began researching through various mathematical sources, trying to find proofs behind James' models. We were able to find an article titled *Bayesball: A Hierarchical Model for Evaluating Fielding in Major League Baseball*. Fielding is somewhat of an overlooked aspect of the game. Sure, defense is extremely important, but when defense is studied or talked about, the fo-

cus usually quickly turns to the pitcher. Fielding percentages[1] do not have much variation across Major League Baseball, since errors among players at that level are not common. Moreover, it is not very telling about the actual skill of a defender. For example, suppose that there are two players, player 1 and player 2, and they both - in seperate instances - attempt the same ball in play hit at the same speed in the same direction, and assume it is a relatively difficult play to make. Now suppose player 1 has enough range and awareness to get to the ball, but the ball hits off of his glove and he does not successfully field the ball. Now assume that player 2 does not have enough range and awareness to get to the ball quick enough, and it simply results in a base hit. In this instance, player 1 would be charged with an error, thereby lowering his fielding percentage. However, player 2 would not be charged with an error, since result would technically not be considered an error, and would not be expected to make that play on a regular basis. As demonstrated by this example, we can see that fielding percentage does not provide enough information about a player's defensive skill to adequately and completely evaluate fielders.

For this project, we will exclusively study data with respect to **position players**. Baseball terminology implies the distinction between pitchers and position players. For this reason, and the fact that pitchers also very rarely field balls in play that are meaningful to this project, we omit pitchers from this project. Even though baseball terminology implies the consideration of the catcher as a position player, the catcher is also omitted from these models. The catcher very seldom fields a ball in play, and if he does, it is usually either a very poorly-hit ball, or a bunt - both of which would only weaken our models if included. Therefore, catchers are omitted altogether for the sake of this project as well. Consequently, the positions of players in this project include all infielders and outfielders - the infielders will include the third baseman, shortstop, second baseman, and first baseman, and the outfielders will include the left fielder, center fielder, and right fielder.

[1] A player's fielding percentage equals the number of successfully fielded balls in play divided by the total number of attempts at a ball in play the player has. A failed attempt at fielding a batted ball is known as an error, and errors lower a players fielding percentage. A pitcher's fielding percentage is not frequently studied, if at all, which explains the separation between pitchers and other fielders that is implied in this instance.

Each position on the field is designated a certain zone on the field. For each position player, there is what can be called a standard position. It is a rule of baseball that the pitcher must stand on the mound in order to legally make each pitch, so even though the pitcher is omitted from this project, he would have to be omitted from describing a shift anyway. A defensive shift in baseball can be defined as a deviation of one or more fielders from their respective standard positions on the field. The general reason for a shift is to increase fielders' chances of successfully fielding a BIP, given certain tendancies of specific batters. Are defensive shifts always worth it, though? This project intends to quantify the success rates of fielders, particularly those that are shifited from standard position. Moreover, it will observe how the success of fielders would have changed if they had been in standard position, rather than shifted. That change in success can conseuqently have larger affects on overall team success, and ultimately could affect number of team wins in the long run.

2
Modeling the Fielders

2.1 The Probabilistic Model of Success

This model is a representation of the probability that player i will successfully field a particular ball in play j - hereafter referred to as BIP j. As the adjustment for this project calls for, however, it will actually model the probability that a player playing a certain position will successfully field BIP j. This probability, denoted p_{ij}, is not uniform across all players for all BIP. We must make slight distinctions between p_{ij} with respect to fly balls/liners and ground balls, since there are different metrics and different players involved. We will specifically make these distinctions in the following subsections.

2.1.1 Fly Balls/Liners

The probability function for fly balls/liners requires that we use a two-dimensional spatial representation, since the distance that player i will travel across a two-dimensional (x, y) plane to catcn BIP j[5]. For a fly ball and a liner, the method for calculating p_{ij} will be the same. All players are considered eligible to field a fly ball or liner (a fly ball to an infielder is considered a pop-up). Then, the following formula represents the probability that player i successfully fields BIP j, such that BIP j is a fly ball or liner:

$$p_{ij} = \Phi(\beta_{i0} + \beta_{i1}D_{ij} + \beta_{i2}F_{ij} + \beta_{i3}V_{ij}) \tag{2.1.1}$$

such that D_{ij} denotes the distance player i needed to travel to field BIP j, F_{ij} denotes whether player i needed to move forward ($F = 1$) or backward ($F = 0$), and V_{ij} denotes the velocity at which BIP j is traveling when it is hit. We will explain these variables and their calculations in Subsection 2.2.2.

In this case, $\Phi(\cdot)$ denotes the cumulative distribution function for the standard Normal distribution[5]. Here, p_{ij} is a function of D_{ij}, F_{ij} and V_{ij}. We notice that more than one beta value is used. This means that different betas are considered for each type of situation of which p_{ij} is a function. The β_{i0} parameter controls the situation in which player i does not need to move to field the BIP - in other words, it is hit directly at him, or $D_{ij} = 0$. The β_{i1}, β_{i2}, and β_{i3} parameters measure the range of player i, whether player i moved forward or backward, and how the velocity at which BIP j was hit affects player i's ability to catch the BIP, respectively[5].

As briefly menioned earlier, we are treating these variables as a draw from a cumulative standard Normal distribution. The rationale behind doing this is that we want all of our variables to be calculated with respect to the same scale. Our data sets, as will be more thoroughly explained later, are not all on the same scale, as they have different sample sizes, etc. Therefore, we observe these probabilities as denoted by $P(Y = 1|X)$, such that the event Y represents player i successfully fielding a BIP - in which case Y would take the value 1 in the event of success - and X denoting the event that BIP j is hit to player i with certain characteristics while player i is standing in a certain starting position. We have that the success of player i is represented by $Y = X^T\beta$, such that the event Y is a function of each variable crossed with their corresponding β coefficients. Now suppose that there is some auxilary variable $Y^* = X^T\beta + \epsilon$, such that $\epsilon \sim N(0, 1)$. Then Y can be observed as an indication as to whether or not our new

auxilary variable will be positive. Then,

$$\begin{aligned} P(Y = 1|X) &= P(Y^* > 0) \\ &= P(X^T\beta + \epsilon > 0) \\ &= P(\epsilon > -X^T\beta) \\ &= P(\epsilon < X^T\beta) \\ &= \Phi(X^T\beta) \end{aligned}$$

[2]. Using the standard Normal distribution allows for less variability with respect to the data, which allows for our results to be observed with respect to the same scale, as more thoroughly explained later.

2.1.2 Ground Balls

Given the circumstances, time, and resources alotted to this project, we actually were not able to execute these models for ground balls. However, this is a portion of this project's research that would greatly enhance and expand its results. Therefore, this portion of the project has to be left for future work, but we will briefly explain it here nonetheless.

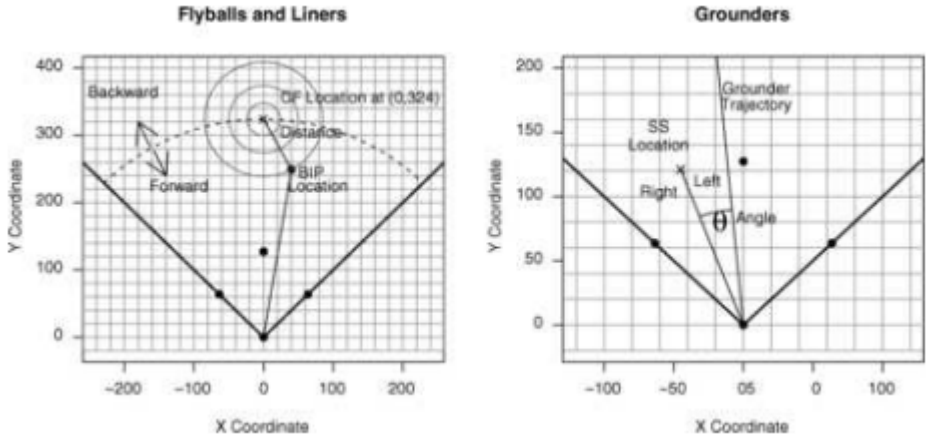
The probability function for ground balls requires us to use a one-dimensional spatial representation, since player i travels in the one-dimensional path of an arc to field BIPs that are ground balls[5]. Unlike in the function for fly balls/liners, not every position is considered eligible for a ground ball. Even though outfielders do field ground balls to the outfield on occasion, a failure to field this type of BIP is tremendously rare. Additionally, a grounder to the outfield is conceptually different than that to an infielder, given that this situation essentially always is the consequence of a routine base hit by the batter. Therefore, we omit outfielders from being eligible for ground balls, leaving infielders to be the only eligible fielders for this type of BIP.

We denote a the angle between player i and the location of the BIP as θ_{ij}. Similarly, the β_{i0} parameter controls the situation in which player i does not need to move to field the BIP - in other words, it is hit directly at him, or $\theta_{ij} = 0$. The β_{i1} and β_{i2} parameters measure the range of player i, or the situations in which he moved to his right and left, respectively. The β_{i3} and

β_{i4} adjust the probability with respect to the velocity of the BIP. Then, the following formula represents the probability that player i successfully fields BIP j, such that BIP j is a ground ball:

$$p_{ij} = \Phi(\beta_{i0} + \beta_{i1}\theta_{ij} + \beta_{i2}L_{ij} + \beta_{i3}V_{ij}). \tag{2.1.2}$$

As in that of fly balls/liners, $\Phi(\cdot)$ denotes the cumulative distribution function for the standard Normal distribution[5].



(2.1.3)

Figure 2.1.3 above depicts the different variables for flyballs/liners and ground balls, with the (x, y) plane projected onto a general representation of a baseball field[5]. The (x, y) plane depicted here, however, is not actually the one that we ultimately have for our data sets. This is simply the one that was used in the *Bayesball* article, and moreover this figure should be observed for the sole purpose of visual representation of all the independent variables within our models.

2.1.3 Accounting for the Batter

We have established the probability of success with respect to fielding, but we also must consider the probability that a batter will actually hit the ball there. Just as no two fielders can be mathematically treated the same, a similar principle applies to batters as well. Since every batter is different, there are different probabilities that a given batter will hit the ball to a

particular location. Consequently, we must take into account the probability that a BIP will be hit at a particular location.

This data was acquired from the *Fangraphs.com* database, and its aquisition was quite simple. I generated a custom report for a certain group batters, which will be explained later, and the database automatically returns numerous statistics for those players. Among these is the percentages of BIP that each batter hit the ball to left field, center field, and right field. The probability that a batter will hit the ball in a particular location can simply be calculated by dividing the number of times the batter hit a BIP in that location by the total number of BIP hit by that batter in a given season. There are a number of extraneous factors that arise from measuring this, which forces us to simplify the measure to be calculated as described. For example, the direction a batter hits a BIP depends on numerous other factors, such as the pitcher's tendencies, the count, the game situation, what part of the season it is, etc. Since we cannot quantify all of the factors, the probability calculations must be simplified to the quotient

$$\%\ \text{BIP}_L = \frac{\#\ \text{BIP}_L}{\text{Total}\ \#\ \text{BIP}} \tag{2.1.4}$$

such that the subscript L denotes the location to which that particular BIP was hit. This data need not to have been manually calculated ourselves, since the *Fangraphs.com* database provided us this information.

The way we will employ this in our model will be through the use of conditional probabilities. The **conditional probability** of an event A on an event B, denoted $\text{P}(A|B)$, can be written as

$$\text{P}(A|B) = \frac{\text{P}(AB)}{\text{P}(B)}. \tag{2.1.5}$$

[6]. This formula is slightly adjusted for this project. Take the center fielder for example. To find the total probability of the center fielder successfully fielding a BIP, we must consider the probability of success by the center fielder, as modeled by p_{ij}, on the condition that a BIP was hit there. Thus, the total conditional probability must take into account the probability of success by the center fielder, on the condition that he is standing at a certain position when the

BIP is hit. Consequently, the total conditional probability can be modeled as follows:

$$P(A) = P(A|B_0)P(B_0) + P(A|B_1)P(B_1) + P(A|B_2)P(B_2) \tag{2.1.6}$$

where event A denotes that in which the center fielder successfully fields a ball, event B denotes the event in which a BIP was hit to either of the three outfield positions, and B_0, B_1, and B_2 denote the events in which the BIP was hit to center field, left field, and right field, respectively[6]. In other words, this allows us to observe the total probability that a center fielder, in this example, will catch any BIP hit to the outfield. This formula is used for each of the three outfield positions.

It may seem a bit counterintuitive to consider the probability that a center fielder catches a BIP that is hit to left field because, clearly, one would have to assume that the left fielder would catch that particular BIP. Therefore, the probability that the center fielder catches this BIP would be much lower. This is true because the fact that the center fielder would be traveling a much further distance to field this BIP than the left fielder would, making the catch more difficult for the center fielder. Moreover, the D_{ij} for the center fielder in this case are larger than those of the left fielder, so the model can aptly demonstrate this increased difficulty for the center fielder in this example.

2.1.4 The Probit Regression for Beta Values

Specifically to help calculate the beta values for individual player evaluation, we can utilize R to run a Probit regression. For this project, however, we will actually pool all the players from each position into one consideration for each position, giving us beta calculations that represent an "average" player at each position. This method is explained furtehr in Subsection 2.2.2.

The Probit model is a statistical regression model that observes the relationship between many independent variables and what effect they have on a dependent variable, such that the dependent variable can only take two values. In this case, the two values for the player evaluation would be fielding a given BIP successfully or not successfully, namely 1 or 0, respectively. To learn how to execute this regression in R, I used examples and from a UCLA online source to help

understand the R code and syntax [3]. I then typed the example code into R, troubleshooting as the code progressed.

2.2 Data Used in the Models

2.2.1 Data Collection

The data used for this entire project was aquired both from the Baseball Savant database of *MLB.com*, and from the databases of *Fangraphs.com*. These are both very renowned and reputable sources, and are used by baseball statisticians from various different careers and fields.

The Baseball Savant database allows the user to search their entire database for specified situational indicators. These indicators include, but are not limited to, to which position BIP were hit, how hard BIP were hit, with how many runners on base BIP were hit, the time period you would like to observe, and even which specific batters you would like to exclusively include in the data set. The data set used for this project, in general terms, included batted balls that were hit to each specified location, from a certain group of batters, throughout the entire 2016 MLB regular season.

This certain group of 42 MLB batters consist of the 42 most shifted-on batters in the 2016 season. In other words, it includes the batters who had the 42 highest number of plate appearances against any type of shift. 42 batters seems like an arbitrary number, and it is. However, I chose 42 since this is a project about baseball, and 42 is the only jersey number that is retired throughout the entire MLB. In other words, 42 can never be worn by any player ever again, because it was the number of Jackie Robinson, who broke MLB's color barrier in 1947. This was determined from *Fangraphs.com*, in which I was able to search for statistics in the 2016 season against all shifts, and I sorted the results from this search by number of plate appearances.

Once those players were determined, I conducted a custom search with respect to those players using the Baseball Savant database. Since shifts are almost always used when there are no runners on base, I created a report of all those players' BIP, such that they were during the 2016 regular

season, no runners were on base, the BIP were either hits or outs[1], and the BIP were hit to either center field, left field, right field, second base, or shortstop. These positions are those that are most affected when a defensive shift is implemented, so we simplified the data set to include only these positions.

Once this data set - hereafter referred to as Set 1 - was created, I exported the data into a Microsoft Excel spreadsheet, and the calculations and mathematical work officially began.

2.2.2 The Data

The models are based on real life statistical data. They will be generally based on the outcome of the individual situation in which ball in play j - hereafter referred to as BIP j - is hit to player i. There are different measurements both with respect to the BIP and to the player. We will also differentiate between two main types of BIP, namely fly balls/liners, and ground balls[2]. Moreover, for the sake of this project, we will be treating fly balls and liners as the same type of BIP, since we intend to only differentiate between balls hit in the air (regardless of their relative height), and balls hit on the ground.

For BIP j, the measurements include the location at which it landed or was fielded, and the velocity of the BIP. For fly balls and liners, the place at which the BIP landed or was caught is based on an (x, y) location on the field, with respect to a Euclidian plane in 2-space representing the field. This was given in Set 1, so for each individual BIP, Set 1 recorded the (x, y) coordinate at which the BIP was hit. For ground balls, the location of the BIP is determined by the angle between the BIP and the x-axis. The velocity of BIP j fielded by player i, denoted V_{ij}, has the same measurement for fly balls and liners and for ground balls, and is subdivided into three categories: $V_{ij} = \{\text{soft, medium, hard}\} = \{1, 2, 3\}$, respectively. A different way of measuring V_{ij} is by actually measuring the velocity in miles-per-hour of each BIP j. However, considering whether a BIP was hit soft, medium, or hard can actually be somewhat more effective when

[1]The hits included all types of hits except for home runs and ground rule doubles, and the outs included flyballs, lineouts, and ground balls.

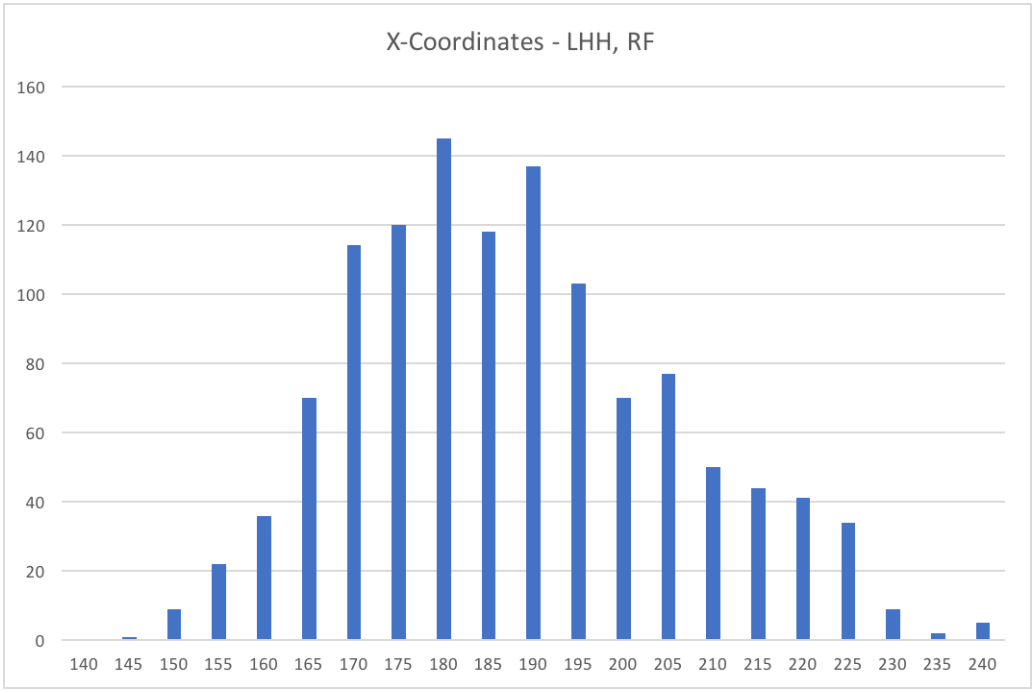
[2]a fly ball is a BIP hit relatively high in the air, and a liner is a BIP that is hit in the air, but at a much lesser height than a fly ball. A ground ball is a BIP that hits the ground one or more times before it is fielded by a defender

interpreting player performance. Moreover, it is the method that was executed by those who built the models, so it followed that we execute such a method as well.

Baseball is often studied with respect to how hard batters hit certain balls. Since humans cannot exactly determine the miles per hour of each batted ball, they must subjectively determine how hard batted balls are hit. For this project, we have made that observation objective. In Set 1, the exit velocity in miles per hour are recorded for each BIP. The average fastball in the MLB today is approximately 92 miles per hour. Suppose that a fastball is thrown at that velocity, and the batter puts the ball in play. If he hits it softly, the BIP velocity will be significantly less than 92 miles per hour, but presumably no lower than about 50 miles per hour - and that would be an *extremely* soft-hit BIP - since the ball must come off the bat at *some* velocity. On the other hand, if the batter squares the ball up and hits it very hard, the exit velocity of the BIP will probably be greater than 92 miles per hour, since the ball was hit hard enough to not only "trampoline" the pitched ball back off the bat, but the batter's strength in his swing will also add a few miles per hour to the velocity of the ball. This is all conceptualized through my extensive knowledge of hitting pitched baseballs, particularly that of the MLB. Moreover, I was able to identify the following parameters for what is to be considered a soft-, medium-, and hard-hit BIP. Any BIP hit less than 75 miles per hour was assigned a 1 for soft, between 75 and 90 miles per hour was assigned a 2 for medium, and above 90 miles per hour was assigned a 3 for hard.

For player i, we will include the distance player i traveled to field BIP j, denoted D_{ij}. The fielders' positions were not given in Set 1, so we had to calculate it ourselves. We had to choose a starting point at which players stood, such that they would be in a position that would reasonably be considered shifted, and such that they each had a meaningful chance to field a BIP hit to their position. Without a clear idea of the $x - y$ plane with which we were working, we had to utilize the data that was available to us that had an indication of the plane - the coordinates of where each BIP landed on the field. To ensure that our chosen starting positions for the fielders would adequately represent those such that the fielders would be able to have

adequate amount of range, and also to ensure that they were in a position of a defensive shift, we decided the best course of action would be to take the average of all the $x - y$ coordinates of each BIP, and consider that x and y coordinate to be the starting position for each fielder. This method is clearly not perfect, but Set 1 is large enough where the average location of a BIP to left field, for example, would actually represent where the left fielder would be standing on any given BIP. Since Set 1 includes the 42 most shifted-on batters in 2016, the calculated coordinates for each position's starting point represents that of a defensive shift, in this instance.



(2.2.1)

An example of the distribution of the x-coordinates of the BIP in Set 1 that were hit to right field, by left-handed hitters. It strongly takes the shape of the normal curve, telling us explicitly that these data points are normally distributed. This histogram was generated for x- and y-coordinates of each BIP location, and by right-handed and left-handed hitters, which served as our check that the coordinates were normally distributed throughout Set 1.

We perform this method for left-handed hitters and right-handed hitters, hereafter referred to as LHH and RHH, respectively. The reason for this is that, as explained earlier, the defensive shift is employed when a batter is assumed to pull the ball the majority of the time. For a LHHs and RHHs, pulling the ball means to hit the ball to the right and left of the CF, respectively.

Therefore, the shifted starting positions of fielders for LHHs and RHHs are slightly different, since they would shift towards different sides of the field for either-handed batter. Therefore, we separate all LHHs and RHHs, for each outfield position to which BIPs are hit. So, we will calculate D_{ij}, using the distance formula: $D_{ij} = \sqrt{(x - x_0)^2 + (y - y_0)^2}$, such that (x, y) is the location of the BIP, and (x_0, y_0) is the starting position for the respective fielder. This was done for outfield shifted starting positions for each outfield position, and for LHHs and RHHs, giving us 6 total starting positions to calculate, and the D_{ij} for BIP to each outfield position are calculated with respect to each of these - leaving us with 3 sets of D_{ij} calculations within Set 1.

We also include the direction player i needed to move to field BIP j. For this measure, we will differentiate it slightly for fly balls/liners and ground balls. For fly balls/liners, we will include whether player i had to move forward or backward, denoted F_{ij}. $F_{ij} = 0$ and $F_{ij} = 1$ if player i had to move backward or forward to field BIP j, respectively. Again, whether the batter was a LHH or RHH matters, since the shifted starting positions of fielders are different for either type of batter. For ground balls, we would include whether player i had to move to his left or right, denoted L_{ij}. $L_{ij} = 0$ and $L_{ij} = 1$ if player i moved to his right or left to field BIP j, respectively. Again, this was not actually executed, since we were limited in resources to move forward with models with respect to ground balls, but we mention it here simply as part of the existing models, which must be left for future work. To calculate this for flyballs/liners, we determined whether the y-coordinate for BIP j was less than that of fielder i's starting position. If it was, then we say player i moved forward, and if not, we say player i moved backward. For groundballs, the same would be done, but with the respective x-coordinates.

Before explaining the last primary metric of the models, suppose there are two situations in which BIP_1 and BIP_2 are hit to player 1 and player 2, respectively, and suppose that both BIP are of the same type, hit to the same place, and at the same velocities. Player 1 and player 2 are obviously not the same player, so their skill levels and skill sets are different. Therefore, it would not follow to assume that the probability that they both field the BIP would be equal.

Consequently, we must implement a metric that represents the skill level of player i. This metric is the beta value, denoted β_i.

This value is calculated through a series of Probit regressions, using fielding data for player i. The data is plugged into this regression using R, and this measure improves our models, because this allows us to treat each outfield position differently, because in reality, we cannot assume that the fielders at each outfield position will behave the same way. Ideally, however, our models could have been more accurate and precise had we calculated our β coefficients for each individual fielder, rather than for all fielders at each outfield position, which is explained further below. We could not do this, however, since we unfortunately did not have the time or resources to take on this workload. Referring back briefly to the discussion about fielding percentage, the β measurements for each outfield position is still a much better metric than solely using fielding percentage as a measure of defensive skill, since, as explained prior, there are known flaws in fielding percentage's ability to provide a reliable basis for evaluating defensive ability.

We had to make a slight adjustment to this methodology as the project progressed. Rather than calculating each β_i for each fielder i, we calculated β values for each position. In other words, we pooled every left fielder, right fielder, and center fielder, and calculated our β values for each of those group of fielders. Therefore, this gave us β coefficients that represented the talent levels of an "average" player at each position.

The method for this was as follows. We had to separate all of our necessary data, namely the V_{ij}, D_{ij}, and F_{ij} for each outfielder, and separated these data sets by position. One more data metric was determined for each individual BIP, namely the Success/Failure metric. In the *Bayesball* paper, the authors consider "eligible" BIP for each fielder. Take the center fielder (CF) for example. They claim that a BIP is "CF-eligible" if it is either caught by the CF, or not caught by any other fielder[5]. For example, a BIP caught by the left fielder (LF) would not be considered "CF-eligible". In addition to those criteria, they further restrict themselves with the notion that any BIP that lands more than 250 feet away from the CF cannot be considered "CF-eligible", since it would not be reasonable to assume a CF should catch a BIP hit more

than 250 feet away from him. In my humble opinion, 250 feet still seems a bit optimistic - I would have set the limit at about 120 feet, which to me seems much more reasonable, as it is hard for myself to envision an outfielder running even 225 feet to field a BIP. In any event, we measured whether each individual BIP was successfully or unsuccessfully fielded, based on whether its result was an out or a hit, respectively. We assigned a 1 for successful, and a 0 for unsuccessful. This gave us the binary variable representing success, which was the last piece we needed before running the probit regression.

When running the probit regression, we needed to run it specifically for each of the three outfield positions, so as to calculate distinct β coefficients for each metric for each position. When running the regression, we call the data from comma-separated values (.csv) files, so we called three separate files - one for each position. Each file contained the data metrics needed to run the regression, and the regression was run, calling the corresponding independent and dependent variables - namely V_{ij}, D_{ij} and F_{ij} as independent variables, and Success/Failure as the dependent variable. The code for the probit regression was as follows:

```
> mydataL <- read.csv("/Users/jackhawke/Dropbox
        (Personal)/Senior_Thesis/savant_probit_data_LF.csv")
> myprobitL <- glm(Success.Failure ~ V_ij + F_ij + D_ij,
        family=binomial(link="probit"), data = mydataL)
> summary(myprobitL)
```

We then model each β coefficient for each position as a draw from a common distribution that is shared by all players at that position[5]. This was done after the β coefficients were calculated, so we did not normalize all the inputs prior. In doing this, this addresses the problem of players not having the same number of events, which consequently can cause unwarranted variability of the β calculation results. Therefore, we modeled each β coefficient as a draw from the standard normal distribution, so as to eliminate this variability.

3
The Shift

3.1 What is a Shift?

Reasons to use the shift vary. The most common reason is to account for a given hitter's tendencies. If a given hitter is known to hit the ball to right field a vast majority of the time, for example, the fielders might determine that a good strategy to give themselves the best chance to field a BIP from this hitter is to move more towards the right field side. This is probably most common in hitters that pull the ball most of the time, which is seen often in hitters that hit more for power[1].

3.2 Affects of the Shift

Since our models are based on real data, shifts that happened in our data set will already be measured as such. However, we can identify when and where a shift occurred, and we can manually adjust these specific situations and create hypothetical scenarios in which a shift did not occur.

[1]There are known to be two main categories of hitting - hitting for power, and hitting for average. Hitters who hit more for power will generally hit a relatively high number of extra base hits and home runs, but consequently will have a slightly lower batting average. Hitter who hit more for average generally hit more singles and doubles and will have slightly higher batting averages.

For all BIP is Set 1, we assume that all fielders are in shifted positions, since it is all data taken from BIP hit by hitters that, with no runners on base, will be extraordinarily likely to be hitting against a shift. Therefore, our shifted positions are those that we calculated for each fielder in Set 1.

3.3 Quantifying the Affect of the Shift

In order to mathematically model the affect that a shift has on fielders' success, we must comparatively observe that success with respect to shifted and non-shifted defensive positioning of those players. By doing this, we can aptly study how effective the shift ultimately is. For example, if player i successfully fielded a BIP while in a shifted position, it would be useful to know the probability of player i fielding the same BIP if he were in a non-shifted, or "standard" position. How much that probability changes, and studying such change, could be the difference between a hit and an out.

Aside from the strictly mathematical side of the game, these probabilities can have a great affect on how fielders are aligned in the future. For example, suppose that player i is in a shifted position, and a BIP is hit right into the shift - in other words, player i was positioned such that the BIP was easily fielded. In this scenario, it is clear that the shift was advantageous for player i, and easily seen as a good strategy for this BIP. However, now suppose that player i were not in that shifted position, but rather in standard position. The question, then, becomes: will player i still successfully field the same BIP, even though he is not shifted? Moreover, how will the likelihood of him doing so change? This is the ultimate question of this project, and how it will be observed is finally explained here.

3.3.1 Adding Non-Shifted Positioning to the Models

In Section 2.2, we discussed the variables that are inputs of the models and how they are calculated, namely D_{ij}, F_{ij}, and V_{ij} as independent variables, and utilizing Success/Failure as the dependent variable to calculate the corresponding β coefficients, consequently multiplying each respective β coefficient with its corresponding independent variable. These ultimately gave

us probabilities that the average fielder of each outfield position will successfully field an arbitrary BIP hit to their position. Moreover, this work was done entirely with respect to Set 1, which contains BIP data on the 42 most shifted-on batters in the 2016 MLB regular season. Therefore, all data is assumed to have been calculated with the understanding that fielders would have been in shifted positions to field all BIP for Set 1.

To study the shifts effectiveness with respect to our models, we cannot use the same data, but rather data that represents outcomes of BIP such that fielders were not in said shifted positions. Since V_{ij} was fixed at 2, we will keep this variable constant for non-shifted data calculations as well. However, D_{ij} and F_{ij} will change, since fielders will be fielding BIP from different starting positions. To calculate starting positions for fielders in Set 1, we took the average of all the (x, y) coordinates of all the BIP, and deemed that to be starting positions for those fielders. Since those were assumed to be shifted positions, we must recalculate the non-shifted starting positions as follows.

Instead of considering the 42 *most* shifted on batters in 2016, we can consider the 42 *least* shifted on batters in 2016. For these batters, since they are shifted on so infrequently, we can assume that fielders would defend these batters in non-shifted, or "standard", positions. Moreover, we can implement the same process for calculating these starting positions, and we did so for this non-shifted set of data. We acquired this data from *Fangraphs.com*, simply by sorting the offensive data for all types of shifts from least to most plate appearances[1]. After find the names of these 42 least shifted on batters, I created a custom report from the Baseball Savant database, using the same process as I did to generate Set 1. I then had a BIP data I would need that represents fielders who are assumed to be in non-shifted positions, hereafter referred to as Set 2.

After Set 2 was generated, the calculations were conducted in the same way as done for Set 1. We determined that calculating average coordinates of the BIP hit to each outfield position to determine the non-shifted starting positions of the fielders at those positions would suffice. Consequently, we calculated D_{ij} and F_{ij} for Set 2, again, the same way as we did for Set 1. With

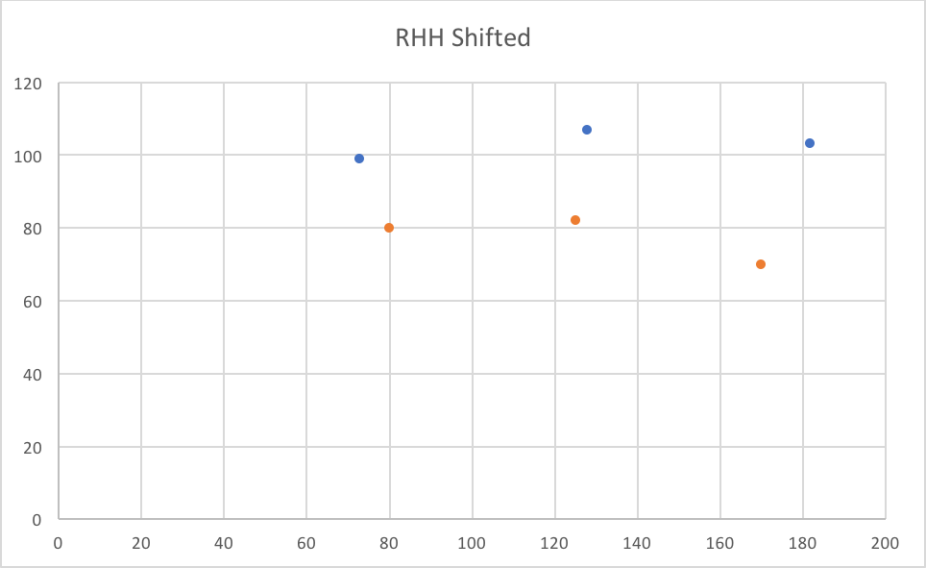
D_{ij} and F_{ij} calculated for each outfield position, and for LHH and RHH, in Set 2, $V_{ij} = 2$ still remaining constant, and Success/Failure being determined for each BIP in Set 2, the necessary variables were the inputs of another probit regression - this time, calculating β coefficients for each corresponding variable, for non-shifted players of Set 2. As done with respect to Set 1, we pool the data from Set 2 for each outfield position, and we then had β coefficients that represented the variables with respect to fielders starting at non-shifted positions.

We now have all we need to proceed to observe the difference in shifted and non-shifted probabilities for fielders at each outfield position.

3.3.2 Calculating the Probabilities

Since we have pooled our β coefficients to represent the talent levels of "average" fielders at each outfield positions, both with respect to shifted and non-shifted starting positions, we must proceed to calculate our respective probabilities based on an assumed BIP that is hit to each of the fielders. Moreover, we needed to have a clear representation of a "typical" BIP hit to each outfield position - one that would likely fall for a base hit. The hit that occurrs most often to any particular outfielder is one that lands in front of them. No research was done to come to this conclusion - in playing and watching baseball my entire life, my intuition strongly convinces me that this would be the most frequent type of hit that occurrs in the outfield. Therefore, we assumed the "average hit" BIP position to be in front of that of the fielders, and we calculated each D_{ij} with respect to each of those. We assumed an average V_{ij}, namely $V = 2$ for each BIP,

and clearly, $F = 1$, since we are assuming that each fielder is moving forward to field the BIP.



(3.3.1)

Above is an example of where the starting locations and the assumed base hits to each fielder are, as plotted on a graph using Microsoft Excel, and such that the fielders' starting locations are shown in blue and the hit locations are shown in orange. If a non-color copy of this project is being viewed, one can observe that there are three "pairs" of points, and each pair has one point with higher y-coordinates than the other. The points of each pair with the greater y-coordinates represent the respective fielders' starting positions, and that with the lesser y-coordinates represent the respective hit locations to those fielders' positions.

To calculate the p_{ij} values, we do so following equation 2.1.1. Once these are calculated, we now have all the variables and metrics needed to carry out the probabiliy calculations. We follow these calculations for LHH and RHH, shifted and non-shifted, and for each outfield position, using equation 2.1.1. Consequently, we finally have our probabilities, of which there are 12 distinct probabilities that are ultimately calculated. They are listed in the table below:

	Shifted		Non-Shifted	
	RHH	LHH	RHH	LHH
Center Fielder	0.79727	0.84550	0.76780	0.76149
Right Fielder	0.08518	0.18981	0.11652	0.18092
Left Fielder	0.22188	0.10937	0.10936	0.10567

(3.3.2)

3.4 Interpretation of Results

The data in Table 3.3.1 can be interpreted as follows. Take the Center Field, Shifted, RHH data point, namely 0.79727. This means that the CF has a 79.727% chance of successfully fielding BIP j hit by a RHH.

This result makes sense, since intuition tells me that a BIP hit to the outfield should be caught by an outfielder between 70% and 80% of the time. The reason for this seemingly arbitrary percentage range is because MLB hitters get hits about 20% to 30% of the time, and if a batter gets a hit 30% or more of the time, he is considered a great hitter. Most hits are to the outfield, but some are infield hits - those in which a batter beats a ground ball throw to first base and is safe, or some other rare mistake occurs in the infield. Therefore, on any BIP that a batter does not get a hit, a fielder should successfully field that BIP (unless a field error is made). My intuition, moreover, tells me that an outfielder should be making successful outs from 70% to 80% of the time, which is where I came up with this percentage range.

Unfortunately, however, the results for Right Field and Left Field are not consistent with this intuition, which they should be. This was a problem that was worked on for weeks - troubleshooting and rerunning the probit regressions, attempting to adjust the starting positions of fielders, etc. - but to no avail could we yield results that fit our intuition. An issue like this can be devastating when conducting a long-term research project like this one, but it seemingly cannot be solved at this point. Fixing this problem is included in tasks I must leave for future work, as outlined in Chapter 4.

This also might have a little bit to do with the fact that we did not normalize all of our data inputs. Had we done that, it would have eliminated a lot of variability in those inputs, such as differences in ballpark dimensions, and would have helped us to observe and calculate with respect to the same scale. Moreover, normalizing our inputs in advance would have allowed subsequent calculations and analyses easier to compare and to observe, since, again, they would all be with respect to the same scale.

Despite the shortcomings of the results themselves, the probabilities relative to each other are actually very telling and interesting. The idea of this project was to measure and observe the change in probabilities of success of these fielders that the defensive shift has. Observing our results relative to each other, we have done that. All but one pair of shifted and non-shifted results reveal that the probability of success for each fielder, and for RHH and LHH, was higher when a shift was used than when a shift was not used. The only pair for which this is not true is the Right Fielder, for RHHs. The reason for this can, at this point, be most aptly identified as randomness with respect to our data sets. Our sample size was so relatively small, and baseball is such a random game in and of itself. Thus, the fact that the Right Fielder would have a better chance of success in non-shifted positioning than in shifted positioning is probably just a random flaw that happened by chance. Again, at this point, there is not much else we can intuitively say about this one pair of results. Nonetheless, our relative results tell us that, in general, employing a defensive shift is advantageous for outfielders.

4
Future Work

4.1 Quantifying the Shift's Effect on Infielders

We followed all the way through with quantifying the effect of the defensive shift on outfielders. We initially intended to do the same for infielders as well, but the primary shortcoming with respect to this was the ambiguity about how to calculate the θ_{ij} values. Since we did not have a clear understanding of how the (x, y) plane was projected onto a field with the Baseball Savant data, it was nearly impossible to calculate θ_{ij}, which is the angle at which a ground ball j was hit to fielder i. Figure 2.1.3 provides a visual representation of θ_{ij}.

Being able to calculate all necessary variables for infielders will allow us to consequently follow through completely with calculating probabilities of success for infielders as well. Therefore, we can have a better understanding of the effect of the shift, not only on outfielders, but on infielders as well.

4.2 Quantifying the Relationship to Wins

Baseball statisticians can quantify how defensive performance affects the number of runs the opposing team scores, which actually can ultimately calculated as one statistical measure. This is known as the Defensive Runs Saved metric. This measurement is based on what is called the

Plus/Minus system, which basically calculates the number of plays a certain defender makes, such that the play made is deviated from an "average" play. In other words, if a player has a Plus/Minus value of, say, 20, that means that he is estimated to make approximately 20 above-average plays in a season.

Consequently, the Plus/Minus values of each player is used when calculating Defensive Runs Saved, since we can say that an above average play in turn prevents a batter from reaching base, which could also ultimately prevent that runner from scoring a run.

Let us look at the plus/minus from a conceptual, estimative standpoint. It is estimated that a little less than half of the plus/minus value represents the approximate runs that player will save[4]. For example, suppose player i is determined to have a plus/minus value of 33 (which is high). Then, conceptually, this player will be estimated to save approximately 15 runs over the course of a season.

Another rule of thumb estimates that the value of a win is approximately that of scoring 10 runs[4]. Therefore, if we divide the estimated number of runs saved by 10, we can have an estimate of the approximate number of wins in a season. Looking at the given example, if player i is estimated to save approximately 15 runs in a given season, he will also be estimated to help his team gain approximately 1.5 wins in a season.

Note that, at this point, I hope to gain a much better understanding of the specific math that goes into these measures, but I have been able to gain a minimal conceptual understanding of how they are calculated. I hope to gain a complete understanding of it as I move forward with my research and the project.

We will be able to quantify how many defensive runs saved, and consequently estimated wins gained, for each player i, given the current positioning that they took in that given season. Then, after adjusting these players out of their defensive shifts, we can recalculate these measures, based on our new findings post-shift adjustment. For example, if we estimate that player i saved approximately 2 wins in a given season, this number will change when we are able to hypothetically take him out of a shift, in situations in which he actually did shift. Since new

defensive runs saved values will presumably be calculated, each player will presumeably affect the number of wins they have helped gain for their team differently, and this will allow us to mathematically estimate how the shift affects win percentage in Major League Baseball.

4.3 Conclusion

Working on this project for an entire academic year was rigorous yet very rewarding. Being able to utilize my extensive baseball knowledge and intuition, while still learning about and making mathematical discoveries and topics was truly enjoyable for me. I hope to be able to add to this project in the future, as I am strongly inclined to continue gathering results that stem from this research and these models. This officially concludes my project, for now at least, and I am very pleased with the way all of this turned out.

Bibliography

[1] *fangraphs.com*, fangraphs.com (2017May).

[2] James H. Albert and Siddhartha Chib, *Bayesian analysis of binary and polychotomous response data*, Journal of the American Statistical Association **88** (1993), no. 422, 669–679.

[3] J. Bruin, *newtest: command to compute new test @ONLINE*, 2011.

[4] John Dewan and Ben Jedlovec, *The fielding bible iv*, Baseball Info Solutions, 2016.

[5] Shane T. Jensen, Kenneth E. Shirley, and Abraham J. Wyner, *Bayesball: A bayesian hierarchical model for evaluating fielding in major league baseball*, Ann. Appl. Stat. **3** (200906), no. 2, 491–520.

[6] Géza Schay, *Introduction to probability with statistical applications*, Birkhäuser Boston, Inc., Boston, MA, 2007. MR2339431